Quantum Package 2.0: An Open-Source Determinant-Driven Suite of Programs

Yann Garniron,† Thomas Applencourt,‡ Kevin Gasperich,‡,§ Anouar Benali,§ Anthony Ferté,¶ Julien Paquier,¶ Barthélemy Pradines,¶,‖ Roland Assaraf,‖ Peter Reinhardt,‖ Julien Toulouse,‖ Pierrette Barbaresco,§ Nicolas Renon,§ Grégoire David,‖ Jean-Paul Malrieu,‖ Mickaël Véril,‖ Michel Caffarel,† Pierre-François Loos,*† Emmanuel Giner,*∥ and Anthony Scemama,*†

†Laboratoire de Chimie et Physique Quantiques (UMR 5626), Université de Toulouse, CNRS, UPS, Toulouse, France
‡Computational Science Division, Argonne National Laboratory, Argonne, Illinois 60439, United States
§Department of Chemistry, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States
¶Laboratoire de Chimie Théorique, Sorbonne Université, CNRS, Paris, France
‖Institut des Sciences du Calcul et des Données, Sorbonne Université, F-75005 Paris, France
#CALMIP, Université de Toulouse, CNRS, INPT, INSA, UPS, UMS 3667, Toulouse, France
‡Aix-Marseille Univ, CNRS, ICR, Marseille, France

ABSTRACT: Quantum chemistry is a discipline which relies heavily on very expensive numerical computations. The scaling of correlated wave function methods lies, in their standard implementation, between \(O(N^3)\) and \(O(\epsilon^N)\), where \(N\) is proportional to the system size. Therefore, performing accurate calculations on chemically meaningful systems requires (i) approximations that can lower the computational scaling and (ii) efficient implementations that take advantage of modern massively parallel architectures. QUANTUM PACKAGE is an open-source programming environment for quantum chemistry specially designed for wave function methods. Its main goal is the development of determinant-driven selected configuration interaction (sCI) methods and multireference second-order perturbation theory (PT2). The determinant-driven framework allows the programmer to include any arbitrary set of determinants in the reference space, hence providing greater methodological freedom. The sCI method implemented in QUANTUM PACKAGE is based on the CIPSI (Configuration Interaction using a Perturbative Selection made Iteratively) algorithm which complements the variational sCI energy with a PT2 correction. Additional external plugins have been recently added to perform calculations with multireference coupled cluster theory and range-separated density-functional theory. All the programs are developed with the IRPF90 code generator, which simplifies collaborative work and the development of new features. QUANTUM PACKAGE strives to allow easy implementation and experimentation of new methods, while making parallel computation as simple and efficient as possible on modern supercomputer architectures. Currently, the code enables, routinely, to realize runs on roughly 2,000 CPU cores, with tens of millions of determinants in the reference space. Moreover, we have been able to push up to 12,288 cores in order to test its parallel efficiency. In the present manuscript, we also introduce some key new developments: (i) a renormalized second-order perturbative correction for efficient extrapolation to the full CI limit and (ii) a stochastic version of the CIPSI selection performed simultaneously to the PT2 calculation at no extra cost.

I. INTRODUCTION

In 1965, Gordon Moore predicted that the number of transistors in an integrated circuit would double about every two years (the so-called Moore’s law). Rapidly, this “law” was interpreted as an expected 2-fold increase in performance every 18 months. This became an industrial goal. The development of today’s most popular electronic structure codes was initiated in the 1990s (or even before). At that time, the increase in computational power from one supercomputer generation to the next was mostly driven by an increase of processors’ frequency. Indeed, the amount of random access memory was small, the time to access data from disk was slow, and the energy consumption of the most powerful computer was 236 kW, hence far from being an economical concern. At the very beginning of the 21st century, having increased continuously, both the number of processors and their frequency raised the supercomputer power consumption by 2 orders of magnitude.
inflating accordingly the electricity bill. The only way to slow this frenetic growth of power consumption while keeping alive Moore’s dream was to freeze the processor’s frequency (between 1 and 4 GHz) and increase the number of CPU cores. The consequence of such a choice was that “free lunch” was over: the programmers now had to parallelize their programs to make them run faster. At the same time, computer scientists realized that the increase in performance in memory access was slower than the increase in computational power and that the floating-point operation (or flop) count would soon stop being the bottleneck. From now on, data movement would be the main concern. This paradigm shift was named the memory wall. Moore’s law is definitely near the end of its life.

The traditional sequential algorithms of quantum chemistry are currently being redesigned and replaced by parallel equivalents by multiple groups around the world. This has obviously a significant influence on methodological developments. The most iconic example of this move toward parallel-friendly methods is the recently developed full configuration interaction quantum Monte Carlo (FCIQMC) method by Alavi and co-workers. FCIQMC can be interpreted as a Monte Carlo equivalent of older selected configuration interaction (sCI) algorithms, such as CIPSI (Configuration Interaction using a Perturbative Selection made Iteratively), that are iterative and thus parallel-friendly methods is the recently developed CIPSI (Con
guration Interaction using a Perturbative Selection made Iteratively),21 that are iterative and thus parallel-friendly methods. In high-precision calculations, the number of determinants that interact with it. To make this step more affordable, we designed a new stochastic scheme which selects on the fly the more important determinants, while the second-order perturbative (PT2) energy is computed using a hybrid stochastic-deterministic scheme. Therefore, the selection part of this new stochastic CIPSI selection is virtually free as long as one is interested in the second-order perturbative correction, which is crucial in many cases in order to obtain near full configuration interaction (FCI) results. Section IV briefly explains how we produce spin-adapted wave functions, and Section V describes parallelism within QUANTUM PACKAGE. The efficiency of the present algorithms is demonstrated in Section VIC, where illustrative calculations and parallel speedups are reported. Finally, Section VII discusses the development of QUANTUM PACKAGE as well as other relevant technical details. Unless otherwise stated, atomic units are used throughout.

II. METHODS

IAA. Generalities. The correlation energy is defined as

$$E_c = E_{\text{exact}} - E_{\text{HF}}$$

(1)

where \(E_{\text{exact}}\) and \(E_{\text{HF}}\) are, respectively, the exact (non-relativistic) energy and the Hartree–Fock (HF) energy in a complete (one-electron) basis set.

To include electron correlation effects, the wave function associated with the \(N\)th electronic state, \(|\Psi_N\rangle\), may be expanded in the set of all possible \(N\)-electron Slater determinants, \(\Pi\), built by placing \(N_1\) spin-up electrons in \(N_{\uparrow}\) orbitals and \(N_\downarrow\) spin-down electrons in \(N_{\downarrow}\) orbitals (where \(N = N_\uparrow + N_\downarrow\)). These so-called molecular orbitals (MOs) are defined as linear combinations of atomic orbitals (AOs)

$$\phi_{\mu}(\mathbf{r}) = \sum_{\nu} C_{\mu\nu} \chi_\nu(\mathbf{r})$$

(2)

Note that the MOs are assumed to be real valued in the context of this work. The eigenvectors of the Hamiltonian \(H\)
are consequently expressed as linear combinations of Slater determinants, i.e.,

$$|\psi_k\rangle = \sum_{i}^{N_{\text{det}}} c_{i|k} |\Psi_i\rangle$$

(3)

where $N_{\text{det}}$ is the number of determinants. For sake of conciseness, we restrict the discussion to the ground state (i.e., $k = 0$) and drop the subscript $k$ accordingly. Solving the eigenvalue problem in this basis is referred to as FCI and yields, for a given basis set, the exact solution of the Schrödinger equation. Unfortunately, FCI is usually computationally intractable because of its exponential scaling with the size of the system.

**II.B. Matrix Elements of the Hamiltonian.** In the $N$-electron basis of Slater determinants, one expects the matrix elements of $H$ to be integrals over $3N$ dimensions. However, given the two-electron nature of the Hamiltonian, and because the MOs are orthonormal, Slater determinants that differ by more than two spinorbitals yield a zero matrix element. The remaining elements can be expressed as sums of integrals over one- or two-electron coordinates, which can be computed at a reasonable cost. These simplifications are known as Slater–Condon’s rules and read

$$\langle il | H | ij \rangle = \sum_{l \in \{l\}} \langle il | h | ij \rangle + \frac{1}{2} \sum_{j \in \{l\}} \langle il | h | ij \rangle$$

(4a)

$$\langle il | H | il \rangle = \langle pl | h | r \rangle + \sum_{l \in \{l\}} \langle pl | r | li \rangle$$

(4b)

$$\langle il | H | pq \rangle = \langle pr | l | qs \rangle$$

(4c)

where $\hat{h}$ is the one-electron part of the Hamiltonian (including kinetic energy and electron–nucleus attraction operators).

$$\langle pl | h | r \rangle = \int f_p(x) \hat{h}(x) f_r(x) dx$$

(5)

are one-electron integrals, $l \in \{l\}$ means that $\phi_l$ belongs to the Slater determinant $|\Psi_l\rangle$ and $|\Psi_p\rangle$ and $|\Psi_q\rangle$ are determinants obtained from $|\Psi_l\rangle$ by substituting orbitals $\phi_p$ by $\phi_l$ and $\phi_q$ by $\phi_l$ and $\phi_q$, respectively.

$$\langle pq | l | rs \rangle = \int f_p(x_l) f_q(x_l) r_{12}^{-1} f_r(x_2) f_s(x_2) dx_1 dx_2$$

(6)

are two-electron electron repulsion integrals (ERIs), $r_{12}^{-1} = r_1 - r_2$ is the Coulomb operator, and $(pq|l|rs) = (pq|rs) - (ps|rq)$ are the usual antisymmetrized two-electron integrals.

Within the HF method, Roothaan’s equations allow us to solve the problem in the AO basis. In this context, one needs to compute the $O(N_{\text{orb}}^4)$ two-electron integrals ($\mu \nu \lambda \sigma$) over the AO basis. Because of a large effort in algorithmic development and implementation, these integrals can now be computed very fast on modern computers. However, with post-HF methods, the computation of the two-electron integrals is a potential bottleneck. Indeed, when computing matrix elements of the Hamiltonian in the basis of Slater determinants, ERIs over MOs are required. Using eq 2, the cost of computing a single integral $(pq|rs)$ scales as $O(N_{\text{orb}}^4)$. A naive computation of all integrals in the AO basis would cost $O(N_{\text{orb}}^8)$. Fortunately, computing all of them can be scaled down to $O(N_{\text{orb}}^3)$ by transforming the indices one by one. This step is known as the four-index integral transformation. In addition to being very costly, this step is hard to parallelize in a distributed way, because it requires multiple collective communications. However, techniques such as density fitting (also called the resolution of the identity), low-rank approximations, or the combination of both are now routinely employed to overcome the computational and storage bottlenecks.

**II.C. Selected CI Methods.** sCI methods rely on the same principle as the usual configuration interaction (CI) approach, except that determinants are not chosen a priori based on occupation or excitation criteria but selected among the entire set of determinants based on their estimated contribution to the FCI wave function. Indeed, it has been noticed long ago that, even inside a predefined subspace of determinants, only a small number of them significantly contributes. Therefore, an on-the-fly selection of determinants is a rather natural idea that has been proposed in the late 1960s by Bender and Davidson as well as Whitten and Hackmeyer. sCI methods are still very much under active development. The main advantage of sCI methods is that no a priori assumption is made on the type of electronic correlation. Therefore, at the price of a brute force calculation, a sCI calculation is less biased by the user’s appreciation of the problem’s complexity.

The approach that we have implemented in Quantum Package is based on the CIPSI algorithm developed by Huron, Rancurel, and Malrieu in 1973 that iteratively selects external determinants $|\alpha\rangle$, determinants which are not present in the (reference or variational) zeroth-order wave function

$$|\psi^{(0)}\rangle = \sum_{l} \langle c_{l|l} |\Psi_l\rangle$$

(7)

at a given iteration — using a perturbative criterion

$$\xi^{(2)} = \frac{\langle \psi^{(0)} | H | \alpha \rangle^2}{E^{(0)} - \langle \alpha | H | \alpha \rangle}$$

(8)

where

$$E^{(0)} = \frac{\langle \psi^{(0)} | H | \psi^{(0)} \rangle}{\langle \psi^{(0)} | \psi^{(0)} \rangle}$$

(9)

is the zeroth-order (variational) energy, and $\xi^{(2)}$ is the (second-order) estimated gain in correlation energy that would be brought by the inclusion of $|\alpha\rangle$. The second-order perturbative correction

$$E^{(2)} = \sum \xi^{(2)} = \sum \frac{\langle \alpha | H | \psi^{(0)} \rangle^2}{E^{(0)} - \langle \alpha | H | \alpha \rangle}$$

(10)

is an estimate of the total missing correlation energy, i.e., $E^{(2)} \approx E_{\text{FCI}} - E^{(0)}$, for large enough expansions.

Let us emphasize that sCI methods can be applied to any determinant space. Although presented here for the FCI space, it can be trivially generalized to a complete active space (CAS) but also to standard CI spaces such as CIS, CISD, or MR-CISD. The only required modification is to set to zero the contributions associated with the determinants which do not belong to the target space.

There is, however, a computational downside to sCI methods. In conventional CI methods, the rule by which determinants are selected is known a priori and, therefore, one can map a particular determinant to some row or column indices. As a consequence, it can be systematically determined to which matrix element of $H$ a two-electron
integral contributes. This allows for the implementation of so-called integral-driven methods that work essentially by iterating over integrals. On the contrary, in (most) sCI methods, the determinants are selected a posteriori, and an explicit list has to be maintained as there is no immediate way to know whether or not a determinant has been selected. Consequently, we must rely on the so-called determinant-driven approach in which iterations are performed over determinants rather than integrals. This can be a lot more expensive since the number of determinants $N_{\text{det}}$ is typically much larger than the number of integrals. The number of determinants scales as $O(N_{\text{orb}}^2)$, while the number of integrals scales (formally) as $O(N_{\text{det}}^2)$. What makes sCI calculations possible in practice is that sCI relies on Monte Carlo CI (MCCI), where determinants are randomly added to the zeroth-order wave function. A more stochastic selection of Monte Carlo CI (MCCI), where determinants are randomly added to the zeroth-order wave function. A more stochastic nature recently developed in Alavi10 is MCCI,93,94 where determinants are randomly added to the zeroth-order wave function. This algorithmically allows for an extremely fast selection of doubly excited determinants by an integral-driven approach.

In order to extrapolate the sCI results to the FCI limit, we have adopted the method recently proposed by Holmes, Umrigar, and Sharma40 in the context of the SHCI method.93,94 It consists of extrapolating the sCI energy, $E^{(0)}$, as a function of the second-order Epstein–Nesbet energy, $E^{(2)}$, which is an estimate of the truncation error in the sCI algorithm, i.e., $E^{(0)} \approx E_{\text{FCI}} - E^{(2)}$.21 When $E^{(2)} = 0$, the FCI limit has effectively been reached. This extrapolation procedure has been shown to be robust, even for challenging chemical situations.9,40–45,34 Below, we propose an improved extrapolation scheme which renormalizes the second-order perturbative correction.

III. IMPLEMENTATION

In this section, we give an overview of the implementation of the various methods present in Q UANTUM PACKAGE. The implementation of the crucial algorithms is explained in detail in the Ph.D. thesis of Garniron10 as well as in the Appendix of the present manuscript.
III. Determinant-Driven Computation of Matrix Elements. For performance sake, it is vital that some basic operations are done efficiently and, notably, the computation of the Hamiltonian matrix elements. This raises some questions about the data structures chosen to represent the two-electron integrals and determinants, as well as their consequences from an algorithmic point of view. This section is going to address these questions by going through the basic concepts of our determinant-driven approach.

IIIA. Storage of Two-Electron Integrals. In QUANTUM PACKAGE, the two-electron integrals are kept in memory because they require a fast random access. Considering the large number of two-electron integrals, a hash table is the natural choice allowing the storage of only nonzero values with a data retrieval in near constant time. However, standard hashing algorithms tend to shuffle data to limit the probability of collisions. Here, we favor data locality using the hash function given in Algorithm 1. This hash function (i) returns the same value for all keys related by permutation symmetry, (ii) keeps some locality in the storage of data, and (iii) can be evaluated in 10 CPU cycles (estimated with MAQAO if the integer divisions by two are replaced by right bit shift instructions.

Algorithm 1: Hash function that maps any orbital quartet \((i, j, k, l)\) related by permutation symmetry to a unique integer.

```plaintext
Function HASH(i, j, k, l): /* Hash function for two-electron integrals */
    Data: i, j, k, l are the orbital indices
    Result: The corresponding hash
    v = min(i, j); 
    r = max(i, j); 
    t = r + r(r - 1)/2; 
    q = min(j, l); 
    s = max(j, l); 
    u = q + s(s - 1)/2; 
    return v + w(w - 1)/2;
```

The hash table is such that each bucket can potentially store \(2^{15}\) consecutive key-value pairs. The 15 least significant bits of the hash value are removed to give the bucket index \([\text{hash} = \text{hash}(i, j, k, l)/2^{15}]\), and only those 15 bits need to be stored in the bucket for the key storage \([\text{hash}(i, j, k, l)\mod 2^{16}]\). Hence, the key storage only requires two bytes per key, and they are sorted in increasing order, enabling a binary search within the bucket. The key search is always fast since the binary search is bounded by 15 misses and the maximum size of the key array is 64 kiB, the typical size of the L1 cache. The efficiency of the integral storage is illustrated in the Appendix A1.

IIIB. Internal Representation of Determinants. Determinants can be conveniently written as a string of creation operators applied to the vacuum state |0\>, e.g., \(a_i^\dagger a_j a_k a_l = |\Omega\rangle\). Because of the Fermionic nature of electrons, a permutation of two contiguous creation operators results in a sign change \(a_i a_j a_k a_l = -a_i a_j a_k a_l\), which makes their ordering relevant, e.g., \(a_i^\dagger a_j a_k a_l = -|\Omega\rangle\). A determinant can be broken down into two pieces of information: (i) a set of creation operators corresponding to the set of occupied spin orbitals in the determinant and (ii) an ordering of the creation operators responsible for the sign of the determinant, known as phase factor. Once an ordering operator \(\hat{O}\) is chosen and applied to all determinants, the phase factor may simply be included in the CI coefficient. Additional information about the internal representation of determinants can be found in Appendix A2.

IIIC. Davidson Diagonalization. Finding the lowest root(s) of the Hamiltonian is a necessary step in CI methods. Standard diagonalization algorithms scale as \(O(N^3_{\text{det}})\) and \(O(N_{\text{det}}^2)\) in terms of computation and storage, respectively. Hence, their cost is prohibitive as \(N_{\text{det}}\) is usually, at least, of the order of a few millions. Fortunately, not all the spectrum of \(\hat{H}\) is required: only the first few lowest eigenstates are of interest. The Davidson diagonalization \((105-109)\) is an iterative algorithm which aims at extracting the first \(N_{\text{states}}\) lowest eigenstates of a large matrix. This algorithm reduces the cost of both the computation and storage to \(O(N_{\text{states}}^2 N_{\text{det}}^2)\) and \(O(N_{\text{states}} N_{\text{det}})\), respectively. It is presented in Algorithm 2, and further details about the present Davidson algorithm implementation are gathered in Appendix A3.

Algorithm 2: Davidson diagonalization procedure. Note that \([.,.]\) stands for column-wise matrix concatenation.

```plaintext
Function DAVIDSON.DIAG (Nstates, U):
    Data: Nstates: Number of requested states
    Data: Ndet: Number of determinants
    Data: U: Guess vectors, \(N_{\text{det}} \times N_{\text{states}}\)
    Result: \(N_{\text{states}}\) lowest eigenvalues eigenvectors of \(\hat{H}\)
    converged ← FALSE;
    while ¬converged do
        Gram-Schmidt orthonormalization of \(U\);
        \(W ← H U\);
        \(h ← U^\dagger W\);
        Diagonalize \(h\): eigenvalues \(E\) and eigenvectors \(y\):
        \(U^* ← U y\);
        \(W^* ← W y\);
        for \(k ← 1, N_{\text{states}}\) do
            for \(i ← 1, N_{\text{det}}\) do
                \(R_k ← w_i^\dagger w_k\);
            end
        end
        converged ← \(\|R\| < \epsilon\);
        U ← \([U, R]\);
    end
    return U;
```

IIID. CIPSI Selection and PT2 Energy. IIID1. The Basic Algorithm. The largest amount of work for this second version of QUANTUM PACKAGE has been devoted to the improvement of the CIPSI algorithm implementation. As briefly described in Section II, this is an iterative selection algorithm, where determinants are added to the reference wave function according to a perturbative criterion.

The \(n\)th CIPSI iteration can be described as follows:

1. The zeroth-order (reference or variational) wave function

\[
|\Psi^{(0)}\rangle = \sum_{i \in I_0} c_i |\Omega\rangle \tag{19}
\]

is defined over a set of determinants \(I_0\), characterized as internal determinants, from which the lowest eigenvector of \(\hat{H}\) are obtained.

2. For all external determinants \(\alpha \notin I_0\) but connected to \(I_0\), i.e., \(\langle \Psi^{(0)} | \hat{H} | \alpha \rangle \neq 0\), we compute the individual
perturbative contribution $\varepsilon^{(2)}_a$ given by eq 8. This set of external determinants is labeled $\mathcal{A}_n$.  
3. Summing the contributions of all the external determinants $\alpha \in \mathcal{A}_n$ gives the second-order perturbative correction provided by eq 10, and the FCI energy can be estimated as $E_{\text{FCI}} \approx E^{(0)} + E^{(2)}$.  
4. We extract $|a^{(2)}\rangle \in \mathcal{A}_n$, the subset of determinants $|a\rangle \in \mathcal{A}_n$ with the largest contributions $\varepsilon^{(2)}_a$, and add them to the variational space $I_{n+1} = I_n \cup \mathcal{A}_n^{(2)}$. In practice, in the case of a single-state calculation, we aim at doubling the size of the reference wave function at each iteration.  
5. Iterate until the desired convergence has been reached.

All the details of our current implementation are reported in Appendix A4. In the remainder of this section, we only discuss the algorithm of our new stochastic CIPSI selection.

III D2. New Stochastic Selection. In the past, CIPSI calculations were only possible in practice thanks to approximations. The first approximation restricts the set $\mathcal{A}_k$ by defining a set of generators. Indeed, it is very unlikely that $|\alpha\rangle$ will be selected if it is not connected to any $|\ell\rangle$ with a large coefficient, so only the determinants with the largest coefficients are generators. A second approximation defines a set of selectors in order to reduce the cost of $\varepsilon^{(2)}_a$ by removing the determinants with the smallest coefficients in the expression of $\Psi^{(0)}$ in $E^{(2)}$. This approximate scheme was introduced in the 1980s and is known as class CIPSI.  

The downside of these approximations is that the calculation is biased and, consequently, does not strictly converge to the FCI limit. Moreover, similar to the initiator approximation in FCIQMC, 8 this scheme suffers from a size-consistency issue. 111 The stochastic selection that we describe in this section (asymptotically) cures this problem, as there is no threshold on the wave function: if the calculation is run long enough, the unbiased FCI solution is obtained.

Recently, some of us developed a hybrid deterministic/stochastic algorithm for the computation of $E^{(2)}$. 112 The main idea is to rewrite the expression of

$$E^{(2)} = \sum_{\alpha} \varepsilon^{(2)}_a \langle \Psi^{(0)} | \hat{H} | \alpha \rangle$$

(20)

into elementary contributions labeled by the determinants of the internal space

$$E^{(2)} = \sum_{\ell} \sum_{a \in \mathcal{A}_k} \varepsilon^{(2)}_a \langle \Psi^{(0)} | \hat{H} | \alpha \rangle = \sum_{\ell} \varepsilon^{(2)}_{\ell}$$

(21)

where

$$\varepsilon^{(2)}_a = \frac{\langle \Psi^{(0)} | \hat{H} | \alpha \rangle}{E^{(0)} - \langle \alpha | \hat{H} | \alpha \rangle}$$

(22)

is the corresponding coefficient estimated via first-order perturbation theory, and $\mathcal{A}_k$ is the subset of determinants $|\alpha\rangle$ connected to $|\ell\rangle$ by $\hat{H}$ such that $|\alpha\rangle \not\in \cup_{k<i} \mathcal{A}_k$. The sum is decomposed into a stochastic and a deterministic contribution

$$E^{(2)} = \sum_{\ell \in \mathcal{D}} \varepsilon^{(2)}_{\ell} + \sum_{k \in \mathcal{S}} \varepsilon^{(2)}_k$$

(23)

where $\mathcal{D}$ and $\mathcal{S}$ are the sets of determinants included in the deterministic and stochastic components, respectively.

The $|\ell\rangle$’s are sorted by decreasing $\varepsilon^{(2)}_{\ell}$ and two processes are used simultaneously to compute the contributions $\varepsilon^{(2)}_k$. The first process is stochastic, and $|\ell\rangle$ is drawn according to $\varepsilon^{(2)}_{\ell}$. When a given $\varepsilon^{(2)}_k$ has been computed once, its contribution is stored such that if $|\ell\rangle$ is drawn again later the contribution does not need to be recomputed. The only update is to increment the number of times it has been drawn for the Monte Carlo statistics. In parallel, a deterministic process is run, forcing us to compute the contribution $\varepsilon^{(2)}_k$ with the smallest index which has yet to be computed. The deterministic component is chosen as the first contiguous set of $\varepsilon^{(2)}_k$. Hence, the computation of $E^{(2)}$ is unbiased, and the exact deterministic value can be obtained in a finite time if the calculation is run long enough. The stochastic part is only a convergence accelerator providing a reliable error bar. The computation of $E^{(2)}$ is run with a default stopping criterion set to $\delta E^{(2)} / E^{(2)} = 0.002$, where $\delta E^{(2)}$ is the statistical error associated with $E^{(2)}$.

We would like to stress that, because of the present semistochastic algorithm, the complete wave function is considered, and that no threshold is required. Consequently, size-consistency will be preserved if a size-consistent perturbation theory is applied.

While performing production runs, we have noticed that the computation of $E^{(2)}$ was faster than the CIPSI selection. Hence, we have slightly modified the routines computing $E^{(2)}$ such that the selection of determinants is performed alongside the computation of $E^{(2)}$. This new on-the-fly CIPSI selection performed during the stochastic PT2 calculation completely removes the conventional (deterministic) selection step, and the determinants are selected with no additional cost. We have observed that, numerically, the curves of the variational energy as a function of $N_{\text{det}}$ obtained with either the deterministic or the stochastic selections are indistinguishable, so that the stochastic algorithm does not harm the selection’s quality.

For the selection of multiple states, one PT2 calculation is run for each state, and as proposed by Angeli et al., 113 the selection criterion is modified as

$$\varepsilon^{(2)}_a = \sum_k \frac{\varepsilon^{(2)}_a}{\max_{\ell} \varepsilon^{(2)}_{\ell}} \langle \Psi^{(0)} | \hat{H} | \alpha \rangle$$

(24)

with

$$\varepsilon^{(2)}_a = \frac{\langle \Psi^{(0)} | \hat{H} | \alpha \rangle}{\langle \Psi^{(0)} | \hat{H} | \alpha \rangle - \langle \alpha | \hat{H} | \alpha \rangle}$$

(25)

This choice gives a balanced selection between states of different multiconfigurational nature.

IV. SPIN-ADAPTED WAVE FUNCTIONS

Determinant-based sCI algorithms generate wave functions expressed in a truncated space of determinants. Obviously, the selection presented in the previous section does not enforce that $\hat{H}$ commutes with $\hat{S}$ in the truncated space. Hence, the eigenstates of $\hat{H}$ are usually not eigenvectors of $\hat{S}$, although the situation improves when the size of the internal space tends to be complete. A natural way to circumvent this problem is to work in the basis of configuration state functions (CSFs), but this representation makes the direct computation of the Hamiltonian less straightforward during the Davidson diagonalization.

Instead, we follow the same path as the MELD and SCI codes 114–116 and identify all the spatial occupation patterns

DOI: 10.1021/acs.jctc.9b00176
associated with the determinants.\textsuperscript{117} We then generate all associated spin-flipped configurations and add to the internal space all the missing determinants. This procedure ensures that $\hat{H}$ commutes with $\hat{S}^z$ in the truncated space, and spin-adapted states are obtained by the diagonalization of $\hat{H}$. In addition, we apply a penalty method in the diagonalization by modifying the Hamiltonian as\textsuperscript{118}

$$\hat{H} = \hat{H} + \gamma (\hat{S}^2 - I(\hat{S}^z)_{\text{target}})^2$$

(26)

where $I$ is the identity matrix, and $\gamma$ is a fixed parameter set to 0.1 by default. This improves the convergence to the desired spin state but also separates degenerate states with different spins, a situation that can potentially occur, for example, with Rydberg states. In the Davidson algorithm, this requires the additional computation of $\hat{S}^2 \hat{U}$, for which the cost is expected to be the same as the cost of $\hat{H} \hat{U}$ (see Algorithm 2). The cost of computing $\hat{H} \hat{U}$ and $\hat{S}^2 \hat{U}$ is mostly due to the search of the connected pairs of determinants, namely, the determinants $|i\rangle$ and $|f\rangle$ for which $\langle i | \hat{H} | f\rangle$ and $\langle i | \hat{S}^2 | f\rangle$ are not zero due to Slater–Condon’s rules. We have modified the function computing $\hat{H} \hat{U}$ so that it also computes $\hat{S}^2 \hat{U}$ at the same time. Hence, the search of connected pairs is done once for both operations, and $\hat{S}^2 \hat{U}$ is obtained with no extra computational cost.

Working with spin-adapted wave functions increases the size of the internal space by a factor usually between 2 and 3, but it is particularly important if one is willing to obtain excited states.\textsuperscript{42–44,54} Therefore, the default in \textsc{quantum package} is to use spin-adapted wave functions.

V. PARALLELISM

In \textsc{quantum package}, multiple parallelism layers are implemented: a fine-grained layer to benefit from shared memory, an intermediate layer to benefit from fast communication within a group of nodes, and a coarse-grained layer to interconnect multiple groups of nodes. Fine-grained parallelism is performed with OpenMP\textsuperscript{119} in almost every single routine. Then, to go beyond a single compute node, \textsc{quantum package} does not use the usual single program/multiple data (SPMD) paradigm. A task-based parallelism framework is implemented with the ZeroMQ library.\textsuperscript{120} The single-node instance runs a compute process as well as a task server process, while helper programs can be spawned asynchronously on different (heterogeneous) machines to run a distributed calculation. The helper programs can connect via ZeroMQ to the task server at any time and contribute to a running calculation. As the ZeroMQ library does not take full advantage of the low latency hardware present in HPC facilities, the helper programs are parallelized also with the message passing interface (MPI) for fast communication among multiple client nodes, typically for fast broadcasting of large data structures.

Hence, we have three layers of parallelism in \textsc{quantum package}: OpenMP, MPI, and ZeroMQ. This allows for an elastic management of resources: a running calculation taking too much time can be dynamically accelerated by plugging in more computing resources and by submitting more jobs in the queue or possibly in the cloud, i.e., outside of the HPC facility. This scheme has the advantage that it is not necessary to wait for all the nodes to be free to start a calculation and hence minimizes the waiting time in the batch queue. It also gives the possibility to use altogether different helper programs. For instance, one could use a specific GPU-accelerated helper program on a GPU node, while CPU-only helpers run on the CPU-only partition of the cluster. It is also possible to write a helper program that helps only one PT2/selection step and then exit, allowing it to gather resources after the PT2/selection has started and freeing them for the following diagonalization step.

The current limitation of \textsc{quantum package} is the memory of the single-node instance. We have not yet considered the possibility to add more compute nodes to increase the available memory, but this can be done by transforming the main program into an MPI program using scattered data structures.

We now describe how the Davidson and PT2/selection steps are parallelized.

VA. Davidson Diagonalization. In the direct Davidson diagonalization method, the computational bottleneck is the matrix product $W = H U$, and only this step needs to be distributed. The calculation is divided into independent tasks where each task builds a unique piece of $W$ containing 40 000 consecutive determinants. Communicating the result of all the tasks scales as $O(N_{\text{tasks}})$ independently of the number of parallel processes. On the other hand, $U$ needs to be broadcast efficiently at the beginning of the calculation to each slave process.

The computation of a task is parallelized with OpenMP, looping in a way that guarantees a safe write access to $W$, avoiding the need of a lock. When idle, a slave process requests a task to the ZeroMQ task server, computes the corresponding result, and sends it to the collector thread of the master instance via ZeroMQ. As the OpenMP tasks are not guaranteed to be balanced, we have used a dynamic scheduling, with a chunk size of 64 elements. The reason for this chunk size is to force that multiple threads access to $W$ at memory addresses far apart, avoiding the so-called \textit{false sharing} performance degradation that occurs when multiple threads write simultaneously in the same cache line.\textsuperscript{122} When the task is fully computed, the computed piece of $W$ is sent back to the master process and a new task is requested, until the task queue is empty.

The $U$ and $W$ arrays are shared among threads, as well as all the large constant data needed for the calculation such as the ERIs. Sharing $U$ also provides the benefit to reduce the amount of communication since $U$ needs to be fetched only once for each node, independently of the number of cores. To make the broadcast of $U$ efficient, the slave helper program is parallelized with MPI, and each node runs a single MPI process. The $U$ matrix is fetched from the ZeroMQ server by the process with rank zero, and then, it is broadcast to the other slave processes within the same MPI job via MPI primitives. Then, each MPI process behaves independently and communicates via ZeroMQ with the task server and with the master node which collects the results. A schematic view of the communication is presented in Figure 1.

VB. CIPSI Selection and PT2 Energy. In the computation of $E^{\text{c53}}$ and the CIPSI selection, each task corresponds to the computation of one $\epsilon_i$ or $\epsilon_k$ in eq 23, together with the selection of the associated external determinants. To establish the list of tasks, the Monte Carlo sampling is precomputed on the master node. We associate to each task the number of drawn Monte Carlo samples such that running averages can be computed when the results of the tasks have been received by the collector thread. When the convergence criterion is
As in the Davidson distributed scheme, when the helper processes communicate once from the ZeroMQ server to the rank-zero MPI process, all the common data are broadcast of U starting from the compute process of the master node, gray arrows the exchange of ZeroMQ messages with the task server, and blue arrows the collection of the results.

reached, the task queue is emptied and the collector waits for all the running tasks to terminate.

As opposed to the Davidson implementation where each task is parallelized with OpenMP, here each OpenMP thread handles independently a task computed on a single core. Hence, there are multiple ZeroMQ clients per node, typically one per core, requesting tasks to the task server and sending the results back to the collector thread (Figure 2). Here, all the OpenMP threads are completely independent during the whole selection, and this explains the pleasing scaling properties of our implementation, as shown in Section VIC. As in the Davidson distributed scheme, when the helper programs are run with MPI, all the common data are communicated once from the ZeroMQ server to the rank-zero MPI process. Then, the data are broadcast to all the other processes with MPI primitives (there is one MPI process per node).

VI. RESULTS

VI.A. Capabilities of QUANTUM PACKAGE. Before illustrating the new features of QUANTUM PACKAGE, in the next subsection we propose to give an overview of what can be achieved (in terms of system and basis set sizes) with the current implementation of QUANTUM PACKAGE. To do so, we propose to review some of our very recent studies.

In ref 44, we studied 18 small molecules (water, hydrogen sulfide, ammonia, hydrogen chloride, dinitrogen, carbon monoxide, acetylene, ethylene, formaldehyde, methanamine, thioformaldehyde, acetaldehyde, cyclopropene, diazomethane, formamide, ketene, nitrosomethane, and the smallest streptocyanine) with sizes ranging from one to three nonhydrogen atoms. For such systems, using sCI expansions of several million determinants, we were able to compute more than 100 highly accurate vertical excitation energies with typically augmented triple-ζ basis sets. It allowed us to benchmark a series of 12 state-of-the-art excited-state wave function methods accounting for double and triple excitations.

Even more recently,45 we provided accurate reference excitation energies for transitions involving a substantial amount of double excitation using a series of increasingly large diffuse-containing atomic basis sets. Our set gathered 20 vertical transitions from 14 small- and medium-sized molecules (acrolein, benzene, beryllium atom, butadiene, carbon dimer and trimer, ethylene, formaldehyde, glyoxal, hexatriene, nitrosomethane, nitroxyl, pyrazine, and tetrizene). For the smallest molecules, we were able to obtain well converged excitation energies with an augmented quadruple-ζ basis set, while only augmented double-ζ bases were manageable for the largest systems (such as acrolein, butadiene, hexatriene, and benzene). Note that the largest sCI expansion considered in this study had more than 200 million determinants.

In ref 65, Giner et al. studied even larger systems containing transition metals: [CuCl4]2−, [Cu(NH3)4]2+, and [Cu-(H2O)4]2+. They were able, using large sCI expansions, to understand the physical phenomena that determine the relative energies of three of the lowest electronic states of each of these square-planar copper complexes.

VI.B. Extrapolation. To illustrate the extrapolation procedure described in Section IID, we consider a cyanine dye123 H2N−CH=NH2 (labeled as CN3 in the remaining) in both its ground state and first excited state:45,124,125 The geometry is the equilibrium geometry of the ground state optimized at the PBE0/cc-pVQZ level.125 The ground state is a closed shell, well described by a single reference, while the excited state is singly excited and requires, at least, two determinants to be properly modeled. The calculations were performed in the aug-cc-pVQZ basis set with state-averaged natural orbitals obtained from an initial CIPSI calculation. All the electrons were correlated, so the FCI space which is explored corresponds to a CAS(24,114) space. The reference excitation energy, obtained at the CC3/ANO-L-VQZ level is 7.18 eV124 (see also ref 45). Note that this particular transition is fairly insensitive to the basis set as long as at least one set of diffuse functions is included. For example, we have obtained 7.14 and 7.13 eV at the CC3/aug-cc-pVQZ and CC3/aug-cc-pVTZ levels, respectively.44

In Figure 3, we plot the energy convergence of the ground state (GS) and the excited state (ES) as a function of the number of determinants Ndet with and without the second-order perturbative contribution. From the data gathered in Table 1, one can see that, although E(0) is still large (roughly 0.02 au), the sCl+PT2 and sCl+rPT2 excitation energies converge to a value of 7.20 eV compatible with the reference energy obtained in a larger basis set. We have also plotted the sCl+rPT2 energy given by E(0) + ZE(2) (Section IID2), and we clearly see that this quantity converges much faster than the
usual sCI+PT2 energy. Even for a very small reference wave
function, the energy gap between GS and ES is qualitatively
correct. The graph of Figure 4, which shows the zeroth-order
energy $E^{(0)}$ as a function of the second-order energy $E^{(2)}$
(dotted lines) and its renormalization variant $Z E^{(2)}$ (solid lines),
also indicates that it is practically much easier to extrapolate to
the FCI limit using the rPT2 correction.

As a second test case for rPT2, we consider the widely
studied example of the chromium dimer ($\text{Cr}_2$) in its $^2\Sigma_g^+$
ground state.\cite{10,11,12,13,14} This system is notoriously challeng-
ing as it combines dynamic and static correlation effects, hence
requiring multiconfigurational methods and large basis sets in
order to have a balanced treatment of these two effects. Consequent-
ly, we compute its ground-state energy in the cc-
pVQZ basis set with an internuclear distance
$R_{\text{Cr-Cr}} = 1.68$ Å
close to its experimental equilibrium geometry. Our full-
valence calculation corresponds to an active space CAS-
$28,198$, and the computational protocol is similar to the
previous example. The second-order corrected value $E^{(0)} + E^{(2)}$
as well as its renormalized version $E^{(0)} + ZE^{(2)}$ as a function of

<table>
<thead>
<tr>
<th>$N_{\text{det}}$</th>
<th>$E^{(0)}$ (GS, a.u.)</th>
<th>$E^{(0)}$ (ES, a.u.)</th>
<th>$E^{(0)} + E^{(2)}$ (GS, a.u.)</th>
<th>$E^{(0)} + E^{(2)}$ (ES, a.u.)</th>
<th>$\Delta E$ (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>$-149.499574$</td>
<td>$-149.246268$</td>
<td>$-149.155(1)$</td>
<td>$-149.863(1)$</td>
<td>$7.95(5)$</td>
</tr>
<tr>
<td>58</td>
<td>$-149.519008$</td>
<td>$-149.261390$</td>
<td>$-149.134(1)$</td>
<td>$-149.853(1)$</td>
<td>$7.67(5)$</td>
</tr>
<tr>
<td>131</td>
<td>$-149.537424$</td>
<td>$-149.277496$</td>
<td>$-149.118(1)$</td>
<td>$-149.842(9)$</td>
<td>$7.52(4)$</td>
</tr>
<tr>
<td>268</td>
<td>$-149.559465$</td>
<td>$-149.298484$</td>
<td>$-149.1035(9)$</td>
<td>$-149.830(9)$</td>
<td>$7.42(4)$</td>
</tr>
<tr>
<td>541</td>
<td>$-149.593434$</td>
<td>$-149.323302$</td>
<td>$-149.0845(8)$</td>
<td>$-149.816(8)$</td>
<td>$7.24(4)$</td>
</tr>
<tr>
<td>1101</td>
<td>$-149.627202$</td>
<td>$-149.354807$</td>
<td>$-149.0683(8)$</td>
<td>$-149.804(5)$</td>
<td>$7.18(3)$</td>
</tr>
<tr>
<td>2207</td>
<td>$-149.663850$</td>
<td>$-149.399522$</td>
<td>$-149.0549(7)$</td>
<td>$-149.787(9)$</td>
<td>$7.26(3)$</td>
</tr>
<tr>
<td>4417</td>
<td>$-149.714222$</td>
<td>$-149.448133$</td>
<td>$-149.0409(6)$</td>
<td>$-149.776(2)$</td>
<td>$7.20(3)$</td>
</tr>
<tr>
<td>8838</td>
<td>$-149.765886$</td>
<td>$-149.496401$</td>
<td>$-149.0296(5)$</td>
<td>$-149.765(5)$</td>
<td>$7.19(2)$</td>
</tr>
<tr>
<td>17 680</td>
<td>$-149.817301$</td>
<td>$-149.545048$</td>
<td>$-149.0239(4)$</td>
<td>$-149.761(4)$</td>
<td>$7.14(2)$</td>
</tr>
<tr>
<td>35 380</td>
<td>$-149.859737$</td>
<td>$-149.587668$</td>
<td>$-149.0216(3)$</td>
<td>$-149.758(3)$</td>
<td>$7.17(1)$</td>
</tr>
<tr>
<td>70 764</td>
<td>$-149.893273$</td>
<td>$-149.623235$</td>
<td>$-149.0207(2)$</td>
<td>$-149.756(3)$</td>
<td>$7.18(1)$</td>
</tr>
<tr>
<td>141 545</td>
<td>$-149.919463$</td>
<td>$-149.650109$</td>
<td>$-149.0214(2)$</td>
<td>$-149.757(2)$</td>
<td>$7.18(8)$</td>
</tr>
<tr>
<td>283 108</td>
<td>$-149.937839$</td>
<td>$-149.669735$</td>
<td>$-149.0224(2)$</td>
<td>$-149.757(2)$</td>
<td>$7.20(6)$</td>
</tr>
<tr>
<td>566 226</td>
<td>$-149.950918$</td>
<td>$-149.681278$</td>
<td>$-149.0233(1)$</td>
<td>$-149.758(1)$</td>
<td>$7.21(6)$</td>
</tr>
<tr>
<td>1 132 520</td>
<td>$-149.960276$</td>
<td>$-149.693053$</td>
<td>$-149.0238(1)$</td>
<td>$-149.758(1)$</td>
<td>$7.21(5)$</td>
</tr>
<tr>
<td>2 264 948</td>
<td>$-149.968203$</td>
<td>$-149.700907$</td>
<td>$-149.0240(1)$</td>
<td>$-149.759(1)$</td>
<td>$7.21(4)$</td>
</tr>
<tr>
<td>4 529 574</td>
<td>$-149.975230$</td>
<td>$-149.708061$</td>
<td>$-149.0245(1)$</td>
<td>$-149.759(1)$</td>
<td>$7.21(4)$</td>
</tr>
<tr>
<td>9 057 914</td>
<td>$-149.981770$</td>
<td>$-149.714526$</td>
<td>$-149.0246(9)$</td>
<td>$-149.759(1)$</td>
<td>$7.20(3)$</td>
</tr>
<tr>
<td>18 110 742</td>
<td>$-149.987928$</td>
<td>$-149.720648$</td>
<td>$-149.0249(7)$</td>
<td>$-149.760(5)$</td>
<td>$7.20(3)$</td>
</tr>
<tr>
<td>36 146 730</td>
<td>$-149.993593$</td>
<td>$-149.726253$</td>
<td>$-149.0252(6)$</td>
<td>$-149.760(5)$</td>
<td>$7.19(3)$</td>
</tr>
</tbody>
</table>

"The excitation energy $\Delta E$ (in eV) is the energy difference between the ground state (GS) and the excited state (ES). The statistical error, corresponding to one standard deviation, is reported in parentheses."
the number of determinants in the reference wave function are reported in Table 2 and depicted in Figure 5. Here also, we observe that rPT2 is clearly a superior extrapolation framework compared to the standard PT2 version as it yields a much straighter extrapolation curve, even in the case of a strongly correlated system such as Cr₂. The renormalization factor Z (eq 17) mitigates strongly the overestimation of the FCI energy for small wave functions by damping the second-order energy E(2). Linear extrapolations of the PT2 and rPT2 energies based on the two largest wave functions yield extrapolated FCI energies of −2087.734 and −2087.738, respectively (Table 2). The difference between these two extrapolated FCI energies provides a qualitative idea of the extrapolation accuracy.

**VIC. Speedup.** In this section, we discuss the parallel efficiency of the algorithms implemented in QUANTUM PACKAGE. The system we chose for these numerical experiments is the benzene molecule C₆H₆ for which we have performed sCI calculations with the 6-31G* basis set. The frozen-core approximation has been applied, and the FCI space that we explore is a CAS(30,90). The measurements were made on GENCI’s Irene supercomputer. Each Irene’s node is a dual-socket Intel(R) Xeon(R) Platinum 8168 CPU@2.70 GHz with 192GiB of RAM, with a total of 48 physical CPU cores. Parallel speedup curves are made to 12 288 cores (i.e., 256 nodes) for (i) a single iteration of the Davidson diagonalization and (ii) the hybrid semistochastic computation of E(2) (which includes the CIPSI selection). The speedup reference corresponds to the single node calculation (48 cores).

First, we measure the time required to perform a single Davidson iteration as a function of the number of CPU cores for the two largest wave functions (N_det = 25 × 10⁶ and 100 × 10⁶). The timings are reported in Table 3, while the parallel speedup curve is represented in Figure 6. The parallel efficiency increases together with N_det as shown in Figure 6. For the largest wave function, a parallel efficiency of 66% is obtained on 192 nodes (i.e., 9216 cores). We note that the speedup reaches a plateau at 3072 cores (64 nodes) for N_det = 25 × 10⁶. For this wave function, there are 625 tasks computing each 40 000 rows of W. When the number of nodes reaches 64, the number of tasks is too small for the load to be balanced between the nodes, and the computational time is limited by the time taken to compute the longest task. The same situation arises for N_det = 100 × 10⁶ with 9408 cores (192 nodes), with 2500 tasks to compute.

Second, we analyze the parallel efficiency of the calculation of E(2) for the sCI wave function with N_det = 25 × 10⁶. The stopping criterion during the calculation of E(2) is given by a relative statistical error below 2 × 10⁻⁵ of the current E(2) value. The speedups are plotted in Figure 6 (Table 3). For 192 nodes, one obtains a parallel efficiency of 89%. The present parallel efficiency is not as good as the one presented in the
The reason behind this is a faster computation of $e^{(2)}$, which reduces the parallel efficiency by increasing the ratio communication/computation.

### VII. DEVELOPING IN QUANTUM PACKAGE

#### VIIA. The QUANTUM PACKAGE Philosophy

QUANTUM PACKAGE is a standalone easy-to-use library for developers. The main goals of QUANTUM PACKAGE are to (i) facilitate the development of new quantum chemistry methods, (ii) minimize the dependency on external programs/libraries, and (iii) encourage the collaborative and educative work through human readable programs. Therefore, from the developer point of view, QUANTUM PACKAGE can be seen as a standalone library containing all important quantities needed to perform quantum chemistry calculations, both involving wave function theory, through the determinant driven algorithms, and DFT methods, because of the presence of a quadrature grid for numerical integrations and basic functionals. These appealing features are made more concrete due to the organization of QUANTUM PACKAGE in terms of core modules and plugins (Section VIIC) together with its programming language (Section VIIB), which naturally creates a very modular environment for the programmer.

Although QUANTUM PACKAGE is able to perform all the required steps from the calculation of the one- and two-electron integrals to the computation of the sCI energy, interfacing QUANTUM PACKAGE at any stage, with other programs is relatively simple. For example, canonical or CASSCF molecular orbitals can be imported from GAMESS, while atomic and/or molecular integrals can be

---

Table 3. Wall-Clock Time (in seconds) To Perform a Single Davidson Iteration and a Second-Order Correction $E^{(2)}$ Calculation (which also includes the CIPSI selection) with an Increasing Number of 48-Core Compute Nodes $N_{nodes}$

<table>
<thead>
<tr>
<th>$N_{nodes}$</th>
<th>Wall-clock time (Davidson for $N_{det} = 10^6$)</th>
<th>Wall-clock time (Davidson for $N_{det} = 25 	imes 10^6$)</th>
<th>Wall-clock time (PT2/selection: $N_{det} = 25 	imes 10^6$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3340</td>
<td>65 915</td>
<td>406 840</td>
</tr>
<tr>
<td>32</td>
<td>142</td>
<td>2168</td>
<td>12 711</td>
</tr>
<tr>
<td>48</td>
<td>109</td>
<td>1497</td>
<td>8515</td>
</tr>
<tr>
<td>64</td>
<td>93</td>
<td>1181</td>
<td>6421</td>
</tr>
<tr>
<td>96</td>
<td>93</td>
<td>834</td>
<td>4323</td>
</tr>
<tr>
<td>128</td>
<td>93</td>
<td>674</td>
<td>3287</td>
</tr>
<tr>
<td>192</td>
<td>96</td>
<td>522</td>
<td>2435</td>
</tr>
<tr>
<td>256</td>
<td>96</td>
<td>519</td>
<td>1996</td>
</tr>
</tbody>
</table>

*The statistical error obtained on $E^{(2)}$, defining the stopping criterion, is $0.17 \times 10^{-3}$ a.u.
read from text files like fcidump. Because of this flexibility, some of us are currently developing plugins for performing sCI calculations for periodic systems.

**VIIB. The IRPF90 Code Generator.** It is not a secret that large scientific codes written in Fortran (or in similar languages) are difficult to maintain. The program’s complexity originates from the interdependencies between the various entities of the code. As the variables are more and more coupled, the programs become more and more difficult to maintain and to debug. To keep a program under control, the programmer has to be aware of all the consequences of any source code modification within all possible execution paths. When the code is large and written by multiple developers, this becomes almost impossible. However, a computer can easily handle such a complexity by taking care of all the dependencies between the variables, in a way similar to how GNU Make handles the dependencies between source files.

IRPF90 is a Fortran code generator. Schematically, the programmer only writes computation kernels, and IRPF90 generates the glue code linking all these kernels together to produce the expected result, handling all relationships between variables. To illustrate in a few words how IRPF90 works, let us consider the simple example which consists of calculating the total energy of a molecular system as the sum of the nuclear repulsion and the electronic energy. The electronic energy is the sum of the kinetic and potential energies, i.e., $E_{\text{ele}} = E_{\text{kin}} + E_{\text{pot}}$.

The production tree associated with the computation of the total energy is shown in Figure 7. Within the IRPF90 framework, the programmer writes a provider for each entity, i.e., a node of the production tree. The provider is a subroutine whose only goal is to compute the value associated with the entity, assuming the values of the entities on which it depends are computed and valid. Hence, when an entity is used somewhere in the program (in a subroutine, a function, or a provider), a call to its provider is inserted in the code before it is used such that the corresponding value is guaranteed to be valid.

**QUANTUM PACKAGE** is a library of providers designed to make the development of new wave function theory and DFT methods simple. Only a few programs using these providers are part of the core modules of **QUANTUM PACKAGE**, such as the sCI module using the CIPSI algorithm or the module containing the semistochastic implementation of the second-order perturbative correction. The main goal of **QUANTUM PACKAGE** is to be used as a library of providers, and programmers are encouraged to develop their own modules using **QUANTUM PACKAGE**.

**VIIC. The Plugin System.** External programmers should not add their contributions by modifying directly **QUANTUM PACKAGE**’s core but by creating their own modules in independent repositories hosted and distributed by themselves. This model gives more freedom to the developers to distribute modules as we do not enforce them to follow any rule. The developers are entirely responsible for their own plugins. This model has the advantage to redirect immediately the users to the right developer for questions, installation problems, bug reports, etc.

**QUANTUM PACKAGE** integrates commands to download external repositories and integrate all the plugins of these repositories into the current installation of **QUANTUM PACKAGE**. External plugins appear exactly as if they were part of **QUANTUM PACKAGE**, and if a plugin is useful for many users, it can be easily integrated in **QUANTUM PACKAGE**’s core after all the coding and documentation standards are respected.

Multiple external plugins were developed by the authors. For instance, one can find a multireference coupled cluster program, interfaces with the quantum Monte Carlo programs QMC = Chem, and CHAMP, an implementation of the shifted-Bk method, a program combining CIPSI with RSDFT, a four-component relativistic RSDFT code, and many others.

In particular, **QUANTUM PACKAGE** also contains the basic tools to use and develop range-separated density-functional theory (RSDFT, see, for example, refs 143 and 144), which allows us to perform multiconfigurational density-functional theory (DFT) calculations within a rigorous mathematical framework. In the core modules of **QUANTUM PACKAGE**, single-determinant approximations of RSDFT are available, which fall into the so-called range-separated hybrid (RSH) approximation. These approaches correct for the wrong long-range behavior of the usual hybrid approximations because of the inclusion of the long-range part of the HF exchange. **QUANTUM PACKAGE** contains all necessary integrals to perform RSDFT calculations, including the long-range interaction integrals and Hartree-exchange-correlation energies and potentials derived from the short-range version of the local-density approximation (LDA) and a short-range generalized-gradient approximation (GGA) based on the Perdew–Burke–Ernzerhof (PBE) functional. All numerical integrals are performed using the standard Becke quadrature grid associated with the improved radial grids of Mura et al. With these tools, more evolved schemes based on RSDFT have been developed, such as an energy correlation functional with multideterminant reference depending on the on-top pair density or a basis set correction. The corresponding source code can be found as external plugins (see, for example, https://gitlab.com/eginer/qp_plugins_eginer).

**VIII. CONCLUSION**

Significant improvements were brought to the second version of **QUANTUM PACKAGE**. Some were single-core optimizations, and others focused on the algorithm adaptation to large-scale parallelism (load balancing in particular). Currently, the code has a parallel efficiency that enables routinely to realize runs on roughly 2000 CPU cores, with tens of millions of determinants in the reference space. Moreover, we have been able to push up to 12 288 cores (256 nodes) on GENCI’s supercomputer Irene. Such a gain in efficiency has and will lead to many more challenging chemical applications. The Davidson diagonalization, which is at the center of sCI and FCI methods, suffers from the impossibility to fully store the Hamiltonian in the memory of a single node. The solution we adopted was to resort to direct methods, i.e., recomputing on the fly the matrix elements at each iteration. While an extremely fast method was already available to detect zero
matrix elements, the former implementation still had to search over the $O(N^2_{det})$ matrix elements for interacting determinant pairs. Now, determinants are split in disjoint sets entirely disconnected from each other. Thus, only a small fraction of the matrix elements need to be explored, and an algorithm with $O(N_{det}^{1.2})$ scaling was proposed. While the parallelization of this method was somewhat challenging due to the extremely unbalanced nature of the elementary tasks, a distributed implementation was realized with satisfying parallel speedups (typically 35 for 50 nodes) with respect to the 48-core single-node reference.

Significant improvements were also realized in the computation of the second-order perturbative correction, $E^{(2)}$. A natural idea was to take into account the tremendous number of tiny contributions via a stochastic Monte Carlo approach. $E^{(2)}$ being itself an approximate quantity used for estimating the FCI energy, its exact value is indeed not required, as long as the value is unbiased and the statistical error is kept under control. Our scheme allows to compute $E^{(2)}$ with a small error bar for a few percent of the cost of the fully deterministic computation.

Similarly, the CIPSI selection is now performed stochastically alongside the PT2 calculation. Therefore, the selection part of the new stochastic CIPSI selection is virtually free as long as one is interested in the second-order perturbative correction.

Finally, efforts have been made to make this software as developer friendly as possible due to a very modular architecture that allows any developer to create his/her own module and to directly benefit from all pre-existing work.

A1. Efficiency of Integral Storage

The efficiency of the storage as a hash table was measured on a dual socket Intel Xeon E5-2680 v2@2.80 GHz processor, taking the water molecule with the cc-pVQZ basis set (115 MOs). The time to access all the integrals was measured by looping over the entire set of ERIs using different loop orderings. The results are given in Table 4, the reference being the storage as a plain four-dimensional array.

<table>
<thead>
<tr>
<th>Access</th>
<th>Array</th>
<th>Hash table</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i,j,k,l$</td>
<td>9.72</td>
<td>125.79</td>
</tr>
<tr>
<td>$i,j,l,k$</td>
<td>9.72</td>
<td>120.64</td>
</tr>
<tr>
<td>$i,k,j,l$</td>
<td>10.29</td>
<td>144.65</td>
</tr>
<tr>
<td>$I,k,j,i$</td>
<td>88.62</td>
<td>125.79</td>
</tr>
<tr>
<td>$I,k,i,j$</td>
<td>88.62</td>
<td>120.64</td>
</tr>
<tr>
<td>Random</td>
<td>170.00</td>
<td>370.00</td>
</tr>
</tbody>
</table>

"The time to generate random numbers (measured as 67 ns/integral) was not counted in the random access results."
This is a necessary step to obtain the \((i, j, k, l)\) indices of the two-electron integral(s) involved in the Hamiltonian matrix element between \(|l\rangle\) and \(|j\rangle\). Then, fetching the values of the integrals can be done quickly using the hash table presented in Section IIIA.

Because the data structure used to store determinants implies an ordering of the MOs, we also need to compute a phase factor. Here, we propose an algorithm to perform efficiently the computation of the phase factor. For a determinant \(|l\rangle\) that is going to be used repeatedly for phase calculations, we introduce a phase mask represented as a bitstring:

\[
P_l[i] = 1 \land \sum_{k=0}^N I[k]
\]

where \(\land\) denotes the and bitwise operation, and \(I[k]\) is the \(k\)th bit of bitstring \(I\), corresponding to the \((k + 1)\)th spinorbital of determinant \(|l\rangle\) \((\text{remember that the orbital indices start at 1 and the bit indices start at 0})\). In other words, the \(i\)th bit of the phase mask is set to 1 if the number of electrons occupying the \(i + 1\) lowest spinorbitals is odd, and 0 otherwise. When an electron of determinant \(|l\rangle\) is excited from orbital \(h\) to \(p\), the associated phase factor is

\[
\begin{align*}
+(-1)^{[p-1][h-1]} \oplus I[p-1] & , \quad \text{if } p > h \\
-(-1)^{[p-1][h-1]} \oplus I[p-1] & , \quad \text{if } p > h
\end{align*}
\]

where \(\oplus\) denotes the exclusive or (xor) operation. So if the phase mask is available, the computation of the phase factor only takes a few CPU cycles. Another important aspect is to create efficiently the phase masks. We propose Algorithm 3, which computes it in a logarithmic time for groups of 64 MOs, taking advantage of the associativity of the exclusive or operator. Indeed, the “for” loop executes six cycles to update the mask for \(2^6 = 64\) MOs.

Algorithm 3: Function that returns a phase mask as a bitstring.

```plaintext
Function PhasemaskOfDet (I):
    Data: I: 64-bit string representation of |I\>
    Result: P: phase mask associated with \(|I\rangle\), as a 64-bit string.
    for \(i \in \{1, \ldots, N\}\) do
        \(r \leftarrow 0\);
        for \(i \leftarrow 0, N_{\text{det}} - 1\) do
            \(P_o[i] \leftarrow L_o[i] \land (L_o[i] \ll 1)\);
            for \(d \leftarrow 0, 5\) do
                \(P_o[i] \leftarrow P_o[i] \land (P_o[i] \ll (1 \ll d))\);
            end
            \(P_o[i] \leftarrow P_o[i] \land r\);
            if \((|I_o[i]| \land 1) = 1\) then
                \(r \leftarrow \neg r\);
            end
        end
    end
    return P;
```

\(|I\rangle = \hat{T}|l\rangle\). This is a necessary step to obtain the \((i, j, k, l)\) indices of the two-electron integral(s) involved in the Hamiltonian matrix element between \(|l\rangle\) and \(|j\rangle\). Then, fetching the values of the integrals can be done quickly using the hash table presented in Section IIIA.

A3. Davidson Diagonalization

Within QUANTUM PACKAGE, the Davidson diagonalization algorithm is implemented in its multistate version. Algorithmically, the expensive part of the Davidson diagonalization is the computation of the matrix product \(\mathbf{H} \mathbf{U}\). As mentioned above (Section II), two determinants \(|l\rangle\) and \(|j\rangle\) are connected via \(\mathbf{H}\) (i.e., \(\langle l | H | j \rangle \neq 0\)) only if they differ by no more than two spinorbitals. Therefore, the number of nonzero elements per row in \(\mathbf{H}\) is equal to the number of single and double excitation operators, namely, \(O(N^2_{\text{det}}(N_{ob} - N_f)^2)\). As \(\mathbf{H}\) is symmetric, the number of nonzero elements per column is identical. This makes \(\mathbf{H}\) very sparse. However, for large basis sets, the whole matrix may still not fit in a single node memory, as the number of nonzero entries to be stored is of the order of \(N_{\text{det}}N^2_{\text{ob}}(N_{ob} - N_f)^2\). One possibility would be to distribute the storage of \(\mathbf{H}\) among multiple compute nodes and use a distributed library such as PBLAS\textsuperscript{154} to perform the matrix-vector operations. Another approach is to use a so-called direct algorithm, where the matrix elements are computed on the fly, and this is the approach we have chosen in QUANTUM PACKAGE. This effectively means iterating over all pairs of determinants \(|l\rangle\) and \(|j\rangle\) connected by \(\mathbf{H}\), and if they are accessing the corresponding integral(s) and computing the phase factor. Even though it is possible to compute the excitation degree between two determinants very efficiently,\textsuperscript{152} the number of such computations scales as \(N_{\text{det}}^2\), which becomes rapidly prohibitively high. To get an efficient determinant-driven implementation, it is mandatory to filter out all pairs of determinants that are not connected by \(\mathbf{H}\) and iterate only over connected pairs. To reach this goal, we have implemented an algorithm similar to the Direct Selected Configuration Interaction Using Strings (DISCIUS) algorithm.\textsuperscript{56}

The determinants of the internal space are reordered in linear time as explained in ref 62, such that the wave function can be expressed as

\[
\Psi^{(0)} = \sum_{I_l} \sum_{I_f} C_{I_l I_f} |I_l I_f\rangle
\]

where we take advantage of the Waller—Hartree double determinant representation.\textsuperscript{104}

Moving along a row or a column of \(\mathbf{C}\) keeps the spin-up or spin-down determinants fixed, respectively. For a given determinant, finding the entire list of same-spin single and double excitations can be performed in \(O(N_{\text{det}}) = O(N_{\text{det}}^2) = O(\sqrt{N_{\text{det}}})\), while finding the opposite-spin double excitations is done via a two-step procedure. First, we look for all the spin-up single excitations. Then, starting from this list of spin-up single excitations, we search for the spin-down single excitation such that the resulting opposite-spin doubly excited determinant belongs to \(\Psi^{(0)}\). Hence, the formal scaling is reduced to \(O(N_{\text{det}}^2)\). It could be further reduced to \(O(N_{\text{det}})\) at the cost of storing the list of all singly and doubly excited determinants for each spin-up and spin-down determinant, but we preferred not to follow this path in order to reduce the memory footprint as much as possible.

A4. CIPSI Selection and PT2 Energy

There are multiple ways to compute the \(c_i^{(2)}\)’s. One way is to loop over pairs of internal determinants \(|l\rangle\) and \(|j\rangle\), generate the list of external determinants \(\{|k\rangle\}\) connecting \(|l\rangle\) and \(|j\rangle\),
and increment the corresponding values $c_i^{(2)}$ stored in a hash table. Using a hash table to store in memory a list of $|\alpha\rangle$'s without duplicates and their contributions $c_i^{(2)}$ is obviously not a reasonable choice since the total number of $|\alpha\rangle$'s scales as $O(N_{\text{det}}N_{\text{orb}}^2 (N_{\text{orb}} - N_r)^2)$. To keep the memory growth in check, we must design a function that can build a stream of unique external determinants, compute their contribution $c_i^{(2)}$, and retain in memory only the few most significant pairs $(|\alpha\rangle, c_i^{(2)})$.

In QUANTUM PACKAGE, we build the stream of unique external determinants as follows. We loop over the list of internal determinants (the generators) sorted by decreasing $c_i$. For each generator $|\alpha\rangle$, we gather all the singly and doubly excited determinants $\{|\alpha\rangle\}$, removing from this set the internal determinants and the determinants connected to any other generator $|\beta\rangle$ such that $J < I$. This guarantees that the $|\alpha\rangle$'s are considered only once, without any additional memory requirement.

For each generator $|\alpha\rangle$, before generating its set of $|\alpha\rangle$'s, we precompute the diagonal of the Fock matrix associated with $|\alpha\rangle$. This enables to compute the diagonal elements $(\alpha|H|\alpha)$ involved in eq 8 for a few flops.\(^{35}\) The computation of \(\langle \Psi^{(0)}|H|\alpha\rangle = \sum_c c_i(H|\alpha)\) is more challenging than the diagonal term since, at first sight, it appears to involve the $N_{\text{det}}$ internal determinants. Fortunately, most of the terms amongst this sum vanish due to Slater–Condon’s rules. Indeed, we know that the terms where $|\beta\rangle$ is more than doubly excited with respect to $|\alpha\rangle$ vanish, and these correspond to the determinants $|\beta\rangle$ which are more than quadruply excited with respect to $|\alpha\rangle$.\(^{35}\) To compute efficiently \(\langle \Psi^{(0)}|H|\alpha\rangle\), for each generator $|\alpha\rangle$, we create a filtered wave function $|\Psi^{(0)}\rangle$ by projecting $|\Psi^{(0)}\rangle$ on a subset $J_\alpha$ of internal determinants $\{|\beta\rangle\}$, where $(|\beta|H|\alpha)$ is possibly nonzero. This yields \(\langle \Psi^{(0)}|H|\alpha\rangle = \langle \Psi^{(0)}|H|\alpha\rangle\), where $|\Psi^{(0)}\rangle$ is a much smaller determinant expansion than $|\Psi^{(0)}\rangle$. In addition, as we have defined the $|\alpha\rangle$'s in such a way that they do not interact with $|\beta\rangle$ when $J < I$, all these $|\beta\rangle$'s can also be excluded from $J_\alpha$. This pruning process yielding to $|\Psi^{(0)}\rangle$ will be referred to as the coarse-grained filtering. A fine-grained filtering of $|\Psi^{(0)}\rangle$ is performed in a second stage to reduce even more the number of determinants, as we explain later.

To make the coarse-grained filtering efficient, we first filter out the determinants that are more than quadruply excited in the spin-up and spin-down sectors separately. Using the representation shown in eq A3, this filtering does not need to run through all the internal determinants and scales as $O(N_{\text{det}}) = O(\sqrt{N_{\text{det}}})$. It is important to notice that, at this stage, the size of $J_\alpha$ is bounded by the number of possible quadruple excitations in both spin sectors and does not scale any more as $O(N_{\text{det}})$. Next, we remove the determinants that are (i) quadruply excited in one spin sector and excited in the other spin sector, (ii) triply excited in one spin sector and more than singly excited in the other spin sector, and (iii) all the determinants that are doubly excited in one spin sector and more than doubly excited in the other spin sector.

The external determinant contributions are computed in batches. A batch $I_{pq}$ is defined by a doubly ionized generator $|I_{pq}\rangle = a_p^\dagger a_q^\dagger |\alpha\rangle$. When a batch is created, the fine-grained filtering step is applied to $J_\alpha$ to produce $J_{pq}$ and $\Psi^{(0)}_{pq}$ such that \(\langle \Psi^{(0)}_{pq}|H|\alpha\rangle = \langle \Psi^{(0)}_{pq}|H|\alpha\rangle\).

Each external determinant produced in the batch $I_{pq}$ is characterized by two indices $r$ and $s$ with $\hat{O}a_r^\dagger a_s^\dagger |I_{pq}\rangle = |I_{pq}\rangle$. The contribution associated with each determinant of a given batch is computed incrementally in a two-dimensional array $A(r, s)$ as follows. A first loop is performed over all the determinants $|\beta\rangle$ belonging to the filtered internal space $J_{pq}$. Comparing $|\beta\rangle$ to $I_{pq}$ allows us to quickly identify if $|\beta\rangle$ will be present in the list of external determinants and consequently tag the corresponding cell $A(r, s)$ as banned. Banned cells will not be considered for the computation of $c_i^{(2)}$, nor the determinant selection, as they correspond to determinants already belonging to the internal space. A second loop over all the $|\beta\rangle$ in $J_{pq}$ is then performed. During this loop, all the $(r, s)$ pairs where $|\beta\rangle$ is connected to $|\beta\rangle$ are generated, and the corresponding cells $A(r, s)$ are incremented with $c_i^{(2)}(J|H||\beta\rangle)$.

Figure 8 shows the number of determinants retained in $\Psi^{(0)}_I$ or $\Psi^{(0)}_{pq}$ after filtering out disconnected determinants of the ground state of the CN3 molecule with $N_{\text{det}} = 935$ S22.

<table>
<thead>
<tr>
<th>AUTHOR INFORMATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corresponding Authors</td>
</tr>
<tr>
<td>*E-mail: <a href="mailto:loos@irsamc.ups-tlse.fr">loos@irsamc.ups-tlse.fr</a> (P.-F.L.).</td>
</tr>
<tr>
<td>*E-mail: <a href="mailto:emmanuel.giner@lct.jussieu.fr">emmanuel.giner@lct.jussieu.fr</a> (E.G.).</td>
</tr>
<tr>
<td>*E-mail: <a href="mailto:scemama@irsamc.ups-tlse.fr">scemama@irsamc.ups-tlse.fr</a> (A.S.).</td>
</tr>
</tbody>
</table>
ORCID
Kevin Gasperich: 0000-0003-2210-5498
Anouar Benali: 0000-0002-2133-0338
Peter Reinhardt: 0000-0001-8644-7468
Jean-Paul Malrieu: 0000-0003-0868-8391
Pierre-François Loos: 0000-0003-0598-7425
Emmanuel Giner: 0000-0002-6206-1103

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank the Centre National de la Recherche Scientifique (CNRS) for funding and Cyrus Umrigar for carefully reading the manuscript. Funding from Projet International de Cooperation Scientifique (PICS08310) is also acknowledged. This work was performed using HPC resources from CALMIP (Toulouse) under allocation 2019-18005 and from GENCI-TGCC (Grant 2018-A0040801738). A.B. was supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, as part of the Computational Materials Sciences Program and Center for Predictive Simulation of Functional Materials. K.G. acknowledges support from Grant CHE1762337 from the U.S. National Science Foundation. QUANTUM PACKAGE is licensed under GNU Affero General Public License (AGPLv3).

REFERENCES

(35) Caffarel, M.; Giner, E.; Scemama, A.; Ramírez-Solís, A. Spin Density Distribution in Open-Shell Transition Metal Systems: A Comparative Post-Hartree–Fock, Density Functional Theory, and


